



**Vision:** MLsys. Previous in HPC and Inference. Future in Network and Cloud Computing

## EDUCATION

**Bachelor of Computer Science** Huazhong University of Science and Technology, GPA: **3.95/4.00** 2020.09 — 2024.06  
**PhD of Computer Science** University of Michigan, advisor: **Ang Chen and Mosharaf Chowdhury** 2024.09 — 20xx.xx


## PUBLICATIONS


**MuxServe: Flexible Multiplexing for Efficient Multiple LLM Serving**, Under Review  *Arxiv (system track) ICML'24*  
• *The 41st International Conference on Machine Learning (system track)*  
• Authors: Jiangfei Duan, **Runyu Lu**, Haojie Duanmu, Xiuhong Li, Xingcheng ZHANG, Dahua Lin, Ion Stoica, Hao Zhang


**White-box Compiler Fuzzing Empowered by Large Language Models**, Under Review  *Arxiv ACM CCS'24*  
• *The 31th ACM Conference on Computer and Communications Security*  
• Authors: Chenyuan Yang, Yinlin Deng, **Runyu Lu**, Jiayi Yao, Jiawei Liu, Reyhaneh Jabbarvand, Lingming Zhang

**Accelerating the Reconstruction of Dynamic Graph with Page Remapping**, Under Review **PVLDB'24**  
• *Proceedings of the VLDB Endowment*  
• Authors: \*Hongru Gao, \***Runyu Lu**, Zhiyuan Shao, Hai Jin  
\* denotes joint first authors



## ACADEMIC EXPERIENCE




**Scheduling the Streaming multiprocessors to accelerate LLM Serving** University of California San Diego   
• Role: Research Intern advised by **Prof. Hao Zhang** **LMSYS Lab**, Aug. 2023 — Present  
• Profiled the bottleneck of current SOTA LLM Serving framework(e.g., vllm, ppl.llm).  
• Improve the GPU SM utilization to accelerate the serving throughput of LLMs.

**WhiteFox: White-box Compiler Fuzzing via LLMs** University of Illinois Urbana-Champaign   
• Research Intern advised by **Prof. Lingming Zhang** **ISE Lab**, June. 2023 — Sept. 2023  
• Test optimization in compilers(LLVM IR) with white-box fuzzing technique by leveraging LLMs  
• Detect 96 bugs of Pytorch, TensorFlow XLA, TensorFlowLite, LLVM based on the optimization source code


**Efficient Paged Dynamic Graph Reconstruction** Huazhong University of Science and Technology   
• Research Intern advised by **Prof. Hai Jin, Prof. Zhiyuan Shao** **CGCL Lab**, Oct. 2022 — June 2023  
• Remap the PageTable of Linux Kernel to accelerate the dynamic graph reconstruction.  
• Speed up existing SOTA algorithms by more than **15x** times.

## INDUSTRIAL EXPERIENCE

**Optimize the LLVM Backend of SenseTime GPU, GPU Compiler** Sensetime , Shanghai.China   
• Role: LLVM Backend Developer April 2023 — Sept.2023  
• Mentor: Wenqiang Yin  
• 4000+ line LLVM GPU Backend Optimization Codes  
• GPU Compiler Optimization and MLIR Triton, Instruction Selection, Instruction Pattern Match, CodeGen Emitter

**Develop High Performance Neural Network Inference Engine** Tencent , Shenzhen.China   
• Role: **Top 20** committer of 302 July 2022 — Nov. 2022  
• Mentor: nihui, with **6k+** followers in Github  
• Optimize high performance neural network operators and math library for NCNN , **18k+** stars in Github, handcraftly optimized for X86/ARM/RISCV/GPU platforms.

## MORE INFO

- If you want to get the papers listed above, please contact me by email(lry89757@gamil.com, runyulu@umich.edu, runyulu@hust.edu.cn). I will response very quickly :)
- For better reading experience and more detailed information, please feel free to visit my  website :)